How Grammatical Features Impact Machine Translation: A New Test Suite for Chinese-English MT Evaluation

Huacheng Song[1,3] **Yi Li**[3], **Yiwen Wu**[2], **Yu Liu**[2], **Jingxia Lin**[2], **Hongzhi Xu**[3]
[1]The Hong Kong Polytechnic University
[2]Nanyang Technological University
[3]Shanghai International Studies University
huacheng.song@connect.polyu.hk, {wu0010en, liuy0243}@e.ntu.edu.sg
jingxialin@ntu.edu.sg, {liyi, hxu}@shisu.edu.cn

Abstract

Machine translation (MT) evaluation has evolved toward a trend of fine-grained granularity, enabling a more precise diagnosis of hidden flaws and weaknesses of MT systems from various perspectives. This paper examines how MT systems are potentially affected by certain grammatical features, offering insights into the challenges these features pose and suggesting possible directions for improvement. We develop a new test suite by extracting 7,848 sentences from a multi-domain Chinese-English parallel corpus. All the Chinese text was further annotated with 43 grammatical features using a semi-automatic method. This test suite was subsequently used to evaluate eight state-of-the-art MT systems according to six different automatic evaluation metrics. The results reveal intriguing patterns of MT performance associated with different domains and various grammatical features, highlighting the test suite's effectiveness. The test suite was made publicly available and it will serve as an important benchmark for evaluating and diagnosing Chinese-English MT systems.

1 Introduction

A test suite or a challenge set is a collection of customized or artificially constructed texts used for exhaustively and systematically diagnosing the hidden faults and specific barriers of models in the field of natural language processing (NLP) (King and Falkedal, 1990; Balkan, 1994). It also comes in handy in machine translation (MT) evaluation and has currently experienced an increased weight in the MT community alongside the significant improvement of average automatic translation quality especially in the era of neural machine translation (NMT) and large language model (LLM) (Burchardt et al., 2017; Kocmi et al., 2023).

By leveraging test suites, it is possible to detect the strengths and weaknesses of apparently perfect MT systems in a linguistically driven fashion and at a fine-grained level. However, on the one hand, most previous studies have concentrated on a limited set of language phenomena (Guillou et al., 2018; Popović, 2019; Mukherjee and Shrivastava, 2023), providing only a narrow view of system capabilities. On the other hand, there is a notable scarcity of research and resources concerning non-Latin script languages, such as Chinese (Chen et al., 2023), which require some special handling of MT systems. These facts underscore the need for a large-scale test suite that covers a broad variety of grammatical features appearing in Chinese-English renderings.

Inspired by the grammatical test suite developed by the German Research Center for Artificial Intelligence (DFKI) (Manakhimova et al., 2023, etc.), we create a test suite for Chinese-English automatic translation focusing on multiple Chinese grammatical features, and based on which we conduct a detailed analysis of the state-of-the-art MT systems, including popular commercial NMT systems and advanced LLMs. The UM parallel corpus in the language pair of Chinese-English (Tian et al., 2014) originally containing segments from seven domains serves as the basis for extracting test sentences for 43 distinct Chinese grammatical features. The final test suite comprises 7,848 well-annotated Chinese sentences (at least 50 items for each grammatical feature), each paired with an English reference translation. We report the performance of eight MT systems and discuss the impact of 43 grammatical features, based on scores generated by six mainstream automatic metrics and supplemented by an analysis of manually identified error cases. We make our test suite, system outputs, evaluation scores, and corresponding codes available online for further research purposes[1].

The main contributions of our work are summarized here: 1) We present a grammatical-feature-

[1]https://github.com/florethsong/testsuite-zh-grammaticalfeature

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 1200–1221
November 15–16, 2024.

based and multi-domain test suite for fine-grained Chinese-English translation evaluation. 2) We perform a linguistically driven evaluation to compare the overall performance of different NMT systems and LLMs. 3) We conduct further analysis of various influencing factors in our study from the aspects of automatic evaluation metrics and other external features of sentences to examine the impacts of different grammatical features.

This paper is structured as follows: Section 2 presents a list of studies that are related to the current work. Section 3 shows the main procedure of the construction of the test suite including data extraction and annotation. In Section 4, we describe the experiments of applying our test suite on the mainstream MT systems and give an analysis of the results. Section 5 provides additional discussions on the other interfering factors that may also interact with grammatical features to impose effects on MT. Section 6 outlines our conclusion and future work.

2 Related Work

In the context of probing linguistically nuanced yet critical weaknesses in MT systems to guide future enhancement, the Conference of Machine Translation (WMT) has introduced test suite tracks since 2018, aimed at receiving in-depth insights into the fine-grained performance of MT systems (Macketanz et al., 2018; Guillou et al., 2018; Rysová et al., 2019; Popović, 2019; Kocmi et al., 2020; Bawden and Sagot, 2023; Mukherjee and Shrivastava, 2023; Manakhimova et al., 2023; Chen et al., 2023, inter alia).

Standing out from many studies of MT evaluations dedicated to one or a few textual factors, e.g. Guillou et al. (2018) on pronouns, Rysová et al. (2019) on discourse-related errors, Popović (2019) on conjunctions, Kocmi et al. (2020) on gender coreference and bias, Bawden and Sagot (2023) on user-generated non-standard content, and Mukherjee and Shrivastava (2023) on multiple domains and writing styles, the series of work by DFKI (Macketanz et al., 2018; Avramidis et al., 2019, 2020; Macketanz et al., 2021, 2022; Manakhimova et al., 2023) constructed a test suite covering more comprehensive linguistic phenomena. This ever-evolving test suite comprises over 10,000 sentences now, covering up to 110 linguistic phenomena, such as false friends, named entities, negations, and so on, and across three translation directions: German $\leftrightarrow$ English, English $\rightarrow$ Russian. By applying the combination of regular expressions and manual checks for annotating the linguistic phenomena, they test the capacity of advanced MT systems submitted to the annual WMT tasks for tackling specific translation difficulties associated with such phenomena. Their latest study (Manakhimova et al., 2023) reveals that the mainstream MT systems face great challenges with certain categories of linguistic phenomena, often in a language-dependent manner. Their detailed findings further enable MT developers to facilitate their systems by considering scenarios prone to failure and then taking corrective actions.

Beyond the translation between alphabetic languages in the Indo-European language family, the task regarding pictographic texts in the Sino-Tibetan language family, represented by Chinese, is also open to exploration. As an analytic and isolating language, Chinese has very different ways of expressing syntactic and semantic relations between constituents, resulting in potential ambiguities that largely rely on context to resolve. The issue becomes more salient in the automatic translation task. Particularly, the presence of certain grammatical features in Chinese will potentially cause different problems. However, the comprehensive exploration and the test suites with attention to various Chinese grammatical features remain largely unconsidered. The only Chinese MT test suite submitted to the WMT was constructed by Chen et al. (2023) for investigating the influence of a limited set of features of Chinese source sentences including words, length, grammar, and entropy. Besides, the studies of Cai and Xiong (2020), Tang et al. (2021), and Song and Xu (2024a,b) provided focused glimpses to some certain Chinese phenomena. They examined the abilities of NMT systems to translate discourse phenomena, negation, and multiword expressions across English and Chinese by using a self-built test suite with annotation of pronouns, discourse connectives, and ellipses, an existing corpus with negation information created by Liu et al. (2018), and an extended dataset of WMT test set, respectively.

Building on the light of DFKI test suites (Manakhimova et al., 2023, etc.) and addressing the lack of Chinese-specific test suites, this study is dedicated to providing an inclusive test suite covering 43 Chinese grammatical features and offering a full evaluation of mainstream MT systems. Addition-

ally, depending on the domain-balanced nature of our basic data, i.e. UM corpus (Tian et al., 2014), our test suite is suitable for comparisons across seven textual domains.

3 Construction of a Test Suite with Chinese Grammatical Features

This section details the processes in test suite construction. We first introduce the theoretical framework of the Chinese grammatical features, then describe the procedures of data selection and annotation, and finally present the result and statistics of the data.

3.1 The Framework of Chinese Grammatical Features

For the categorization framework for Chinese grammatical features, we adopt the one in our previous work (Xu and Lin, 2023), which is developed in accordance with a reference grammar of Chinese (Huang and Shi, 2016). The framework systematically addresses 157 typical linguistic phenomena, i.e. grammatical features, in Chinese, organized across various linguistic aspects, including words, structure, semantics, and pragmatics. Word-level structures are concerned with how a word is formed by morphemes. For instance, reduplication is a typical phenomenon to create new words in Chinese. For example, the adjective 高兴 *gao xing* 'happy' can be reduplicated to form another adjective word 高高兴兴 *gao gao xing xing* 'very happy'. The structure category mainly refers to the syntactic structure of sentences, phrases, and special constructions. This framework identifies three semantic subcategories: semantic roles, aspect, and negation. The pragmatic category includes sentence types, information packaging constructions, attitudinal particles/adverbs, deixis, and anaphora.

Whether a certain grammatical feature that is present in the source language might cause problems in automatic translation is largely dependent on the equivalence of the counterpart phenomenon in the target language. Take reflexives as an example. Both the two languages use reflexive pronouns to denote the antecedent nominal phrase. However, there are also some fine distinctions in their usages, leading to obstacles for cross-lingual translation. As shown in Example (1), while in Chinese the pronoun 你 *ni* 'you' can be optional, the English translation must combine the pronoun 'you' in order to obtain the correct reflexive 'yourself'.

(1) 你要照顾好(你)自己。
ni yao zhaogu hao (ni) ziji
you should take_care good (you) self
'You should take care of yourself.'

Many grammatical features are Chinese-specific, such as classifiers, as shown in example (2), BA constructions as shown in (3), headless NP as shown in (4), and so on. Depending on their grammatical differences to varying degrees, different Chinese grammatical features might impose different effects on MT systems. It is thus necessary to create a test suite that covers various grammatical features with each one associated with a set of examples, which can be used to analyze the effects of different grammatical features on MT systems based on statistical methods.

(2) 一顿晚餐
yi dun wancan
one CLF dinner
'a dinner'

(3) 我把这些书都看完了。
wo ba zhexie shu dou kan wan le
I BA these book all read finish PRF
'I have read all these books.'

(4) 羡慕的是缺乏的。
xianmu de (pro) shi quefa de (pro)
admire DE be lack DE
'What is admired is the lacked.'

3.2 Data Preparation

We extract Chinese source sentences and their corresponding English reference translations from the UM corpus (Tian et al., 2014), a high-quality and large-scale parallel corpus embracing eight distinct domains: Education (abbreviated as 'Edu', with 4.5 million bilingual sentence pairs), Laws ('Laws', 2.2M), News ('News', 4.5M), Science ('Sci', 2.7M), Spoken ('Spk', 2.2M), Subtitles ('Sbt', 3M), Thesis ('Ths', 3M), and Microblog ('Mbg', 5K). We exclude the Microblog section due to its small number of sentence pairs, which is far fewer than the other domains, making it difficult to ensure a rough balance across different domains. We select sentences with 10 to 60 Chinese characters to minimize the impact of source sentences with extreme lengths (excessively long or short) on translation quality as well as to avoid the existence of too many different grammatical features in a single sentence that may mix the effects of them on translations.

Grammatical Feature	Abbreviation	Precision	Agreement	Sum	Edu	Laws	News	Sci	Spk	Sbt	Ths
Verb Phrase	VP	0.83	0.83	220	24	33	32	41	41	27	22
Noun Phrase	NP	0.91	0.98	1440	152	344	190	184	184	145	241
Adjective Phrases	AdjP	0.36	0.44	230	18	46	42	24	19	16	65
Adverb Phrases	AdvP	0.83	0.95	951	109	180	141	152	100	95	174
Pre-verbal Preposition Phrase	PreVPP	0.91	0.94	163	14	29	26	23	27	18	26
Post-verbal Preposition Phrase	PstVPP	0.28	0.77	146	23	20	26	22	29	24	2
Participant Preposition Phrase	PtcpPP	0.89	0.83	301	26	82	44	23	38	27	61
Topic Preposition Phrase	TopPP	0.98	0.98	213	25	30	31	35	31	22	39
Reference Preposition Phrase	RefPP	0.96	0.99	347	42	69	52	49	46	34	55
Condition Preposition Phrase	CondPP	0.51	0.89	105	18	33	18	9	6	3	18
Locative Preposition Phrase	LocPP	0.5	0.93	96	15	18	18	12	13	12	8
Sentence-Initial Preposition Phrase	SentIPP	0.33	0.82	64	7	5	11	9	11	3	18
Space Preposition Phrase	SpcPP	0.76	0.86	155	18	32	20	17	29	24	15
Source Preposition Phrase	SrcPP	0.96	0.96	191	27	29	26	28	26	26	29
Path Preposition Phrase	PathPP	0.8	0.93	132	12	16	23	19	28	16	18
Goal Preposition Phrase	GoalPP	0.65	0.68	127	21	16	22	19	27	19	3
Direction Preposition Phrase	DirPP	0.47	0.48	95	20	6	18	12	19	17	3
Space Extension Preposition Phrase	SpanPP	0.93	0.18	169	25	28	25	22	25	16	28
Standard Classifier	StdCLF	0.98	0.99	195	29	39	28	24	28	22	25
Individual Classifier	IndCLF	0.94	0.97	284	28	58	36	40	41	35	46
Event Classifier	EvCLF	0.97	0.97	184	25	23	24	24	31	27	30
Kind Classifier	KindCLF	0.98	0.99	185	25	23	29	29	27	22	30
Approximation Classifier	ApprCLF	0.35	0.81	68	10	13	14	11	10	6	4
Temporal Sequence Complex Sentence	TmpSCpl	0.99	0.98	176	21	28	29	26	26	19	27
Concessive Complex Sentence	ConcCpl	0.99	0.99	156	20	10	29	27	29	14	27
Causative Complex Sentence	CausCpl	0.46	0.82	82	13	8	17	20	13	2	9
Negation BU	BUNeg	0.96	0.91	222	31	37	30	32	35	23	34
Negation MEI/MEIYOU	MEINeg	0.98	0.97	225	32	33	33	33	36	33	25
Negation in Imperative Sentences	ImpNeg	0.36	0.82	83	8	37	11	12	5	10	0
Sublexical Negation	LexNeg	0.97	0.97	182	23	30	26	28	24	20	31
Negative Polarity Items	NPI	0.98	0.92	166	19	27	28	29	24	14	25
Deixis	Deixis	0.95	0.57	272	37	26	38	46	48	45	32
Reflexive	Refl	0.96	0.76	195	25	26	29	30	31	25	29
Reciprocal	Recp	1	1	174	23	27	26	26	23	20	29
Perfective GUO	GUOPrf	0.84	0.91	154	21	24	27	20	28	26	8
Progressive ZAI	ZAIProg	0.99	0.98	184	26	23	28	27	30	21	29
Passive Construction	Pass	0.66	0.84	131	15	30	23	18	19	17	9
Relative Construction	Rel	0.78	0.85	305	16	119	32	25	31	19	63
Comparative Construction	Cmpr	0.95	0.99	191	25	28	28	27	24	23	36
BA Construction	BA	0.99	0.99	199	21	38	30	28	28	23	31
Copular SHI	SHICop	0.98	0.94	230	34	36	38	31	33	23	35
Verbal LE	VerbLE	0.86	0.96	194	31	10	33	31	28	25	36
Quantifier Only ZHI	ZHIQtf	0.97	0.98	220	22	24	28	25	28	26	28
Overall/Total		**0.81**	**0.87**	**9763**	**1176**	**1793**	**1459**	**1369**	**1379**	**1084**	**1503**
Sentence Number				**7848**	**1035**	**1109**	**1207**	**1187**	**1175**	**927**	**1208**

Table 1: The detailed information of our test suite including the definition of grammatical features and their corresponding numbers of instances in each domain. 'Precision' refers to the precision of our self-created grammatical feature identifier in accordance with the results after manual checking. 'Agreement' shows the inner consistency between the judgments given by the two checkers. The acronyms including 'BU', 'MEI/MEIYOU', 'GUO', 'ZAI', 'SHI', 'LE', 'BA', and 'ZHI', are the specific markers indicating particular grammatical features.

3.3 Grammatical Feature Annotation

In the first step, we use a regular-expression-based tool we previously built in Xu and Lin (2023) to identify the Chinese grammatical features in each source sentence automatically. After all the sentences are annotated with a set of grammatical features, we remove the grammatical features that appear fewer than 30 times in all sentences of each domain to ensure a fairly balanced data distribution in statistics, by which our focus is narrowed to 43 target grammatical features out of 157 in the original framework for our test suite. Then, for each grammatical feature, we randomly select about 210 candidate sentences (30 for each of the seven domains) according to the principle of prioritizing those carrying the fewest labels aiming to reduce the mixed effects of multiple features in one sentence. Since some sentences can finally possess multiple features, certain feature groups may include more than 210 sentences.

In the second step, the automatically generated labels of grammatical features are double-checked by two native speakers well-trained in Chinese linguistics. The screening process is primarily focused on identifying false positive grammatical features assigned to sentences. In cases where annotators disagree, they are required to discuss and reach a final decision together. This filtering process resulted in a validated test suite of a total of 7,848 sentence pairs, including 1,127 pairs with no specific grammatical features. Detailed information on the data is shown in Table 1. We see that the average precision of the automatic annotation tool is about 81% and the agreement of the two annotators in identifying false positives is 87%.

4 Evaluation of Chinese-English MT Systems with the Test Suite

In this section, we use our test suite to evaluate eight popular NMT systems and LLMs with six mainstream automatic metrics. We will briefly outline these systems and metrics, and then describe the results of the experiments we conduct to compare the performance of different MT systems on our whole test suite as well as the subgroups divided by domains and grammatical features.

4.1 Evaluated Translation Systems

Aiming at gaining a broad view of the capabilities of leading MT systems and representative LLMs to tackle diverse Chinese grammatical features, we refer to several widely recognized leaderboards, e.g., WMT (Kocmi et al., 2022), SuperCLUE (Xu et al., 2023), SuperBench[2], and Intento[3]. Eventually, four commercial NMT engines, Baidu, Niu, Google (basic v2 edition), DeepL and four advanced LLMs including Ernie (-4 turbo), Qwen (-turbo), GPT (-4o), and Claude (-3 opus) are selected for evaluation. We apply default settings to the NMTs and conduct a zero-shot translation test for the LLMs, setting the temperatures to 0.01.

4.2 Automatic Evaluation Metrics

We use six automatic metrics, including two string-overlap-based metrics: BLEU (Papineni et al., 2002) and CHRF (Popović, 2015)[4], and four neural-network-based ones: two reference-based metrics: COMET (Rei et al., 2022) and XCMOET (Guerreiro et al., 2023), and two reference-free ones: COMETKIWI-QE (Rei et al., 2023) and XCOMET-QE (Guerreiro et al., 2023)[5].

Based on our observation, different metrics may produce different results in analyzing the effects of various factors impacting MT systems. In the following discussions, we will mainly use the average score of XCOMET and XCOMET-QE (henceforth, X-AVERAGE), which are proven the most accurate metrics conforming to human evaluations by the WMT23 metric shared task (Freitag et al., 2023). We also provide extended discussions about the selection of metrics in Section 5.3. For reference, readers can find the results in all six metrics and their overall average (denoted as AVERAGE) in the Appendices.

4.3 Experimental Results

4.3.1 Comparison of Systems

Comparison of Overall Performance Table 2 shows the overall performance of the eight systems in six different metrics. On average, Google performs the best and Qwen the worst. While most of the metrics give similar ratings, XCOMET and XCOMET-QE slightly favor GPT's performance more than the others. We can also see that NMTs achieve marginally better performance than LLMs across all evaluation results. This is partially attributed to the extremely low scores received by the LLM: Qwen.

Comparison on Domains Table 3 shows the performance of all the MT systems on different domains in X-AVERAGE scores. Generally, all the systems show similar trends across different domains with the highest performance on Spoken and Subtitles and the lowest performance on Thesis and Laws. While it is unquestionable that domain affects automatic translation, the results may also be partially influenced by the distribution of sentence length within each domain. We will give more discussion about the effects of sentence length in Section 5.2. Detailed information about the performance of systems on different domains in all the other metrics can be found in Appendix A. Based on the comparisons between different models, it

[2]The online report of SuperBench by Tsinghua University: https://fm.ai.tsinghua.edu.cn/superbench/#/leaderboard

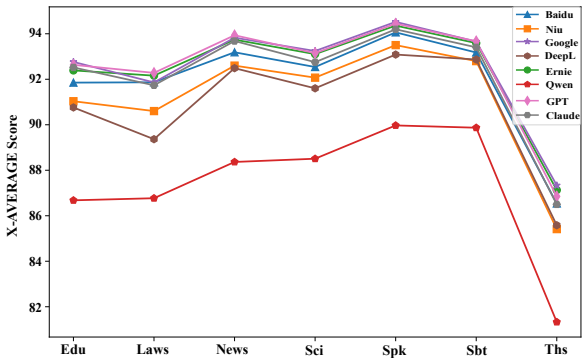
[3]The online report *The State of Machine Translation 2024* by Intento: https://inten.to/machine-translation-report-2024/

[4]BLEU and CHRF are computed by SacreBLEU implementations: https://github.com/mjpost/sacrebleu, with 'True' in signatures of effective order, lowercase, and whitespace and taking 'exp' as smooth method.

[5]The series of *COMET are computed by Unbabel implementations: https://github.com/Unbabel/COMET

	BLEU	CHRF	COMET	XCOMET	COMETKIWI-QE	XCOMET-QE	AVERAGE	X-AVERAGE
Baidu	21.6	56.0	80.8	72.7	90.3	93.4	69.1	91.8
Niu	25.1	58.5	81.0	71.6	89.9	92.3	69.7	91.1
Google	**27.0**	**59.8**	**82.1**	72.8	**91.6**	93.2	**71.1**	**92.4**
DeepL	24.3	57.6	80.4	71.7	89.4	92.1	69.2	90.8
Ernie	24.0	58.1	81.5	73.1	91.3	93.3	70.2	92.3
Qwen	16.5	47.0	76.3	65.0	85.4	89.1	63.2	87.3
GPT	22.5	57.0	81.3	**73.4**	91.1	**93.6**	69.8	**92.4**
Claude	23.3	57.7	81.3	73.0	91.0	93.1	69.9	92.1
NMT-AVG	24.5	58.0	81.1	72.2	90.3	92.8	69.8	91.5
LLM-AVG	21.6	55.0	80.1	71.1	89.7	92.3	68.3	91.0

Table 2: The overall performance of MT systems in different metrics. The highest and the lowest scores among all systems evaluated are highlighted with bold and underlined numbers respectively. 'AVERAGE' is the mean of scores by the six metrics, while 'X-AVERAGE' is the mean of scores by XCOMET and XCOMET-QE.



	Edu	Laws	News	Sci	Spk	Sbt	Ths
Baidu	91.9	91.9	93.2	92.5	94.1	93.2	86.5
Niu	91.0	90.6	92.6	92.1	93.5	92.8	85.4
Google	**92.8**	91.9	93.8	**93.2**	**94.5**	**93.7**	**87.3**
DeepL	90.8	89.4	92.5	91.6	93.1	92.9	85.6
Ernie	92.4	92.2	93.7	93.1	94.4	93.6	87.1
Qwen	86.7	86.8	88.4	88.5	90.0	89.9	81.3
GPT	92.6	**92.3**	**93.9**	93.1	94.4	**93.7**	86.9
Claude	92.5	91.7	93.7	92.8	94.2	93.4	86.5
NMT-AVG	91.6	90.9	93.0	92.3	93.8	93.2	86.2
LLM-AVG	91.1	90.8	92.4	91.9	93.2	92.7	85.4

Table 3: Performance in X-AVERAGE scores of each system on different domains. The highest and the lowest scores among all systems evaluated are highlighted with bold and underlined numbers respectively.

can be pointed out that Google and GPT share the top performance on each domain with a minor difference but Qwen is ranked last across all domains. Again, NMTs show a generally higher performance than LLMs across all domains. However, this is not the case when delving deeper into the specific data excluding Qwen and Google.

Comparison on Grammatical Features Figure 1 shows the performance of all systems in X-AVERAGE scores in each grammatical feature group. The groups are arranged in descending order based on the average scores of all systems. In general, we see that all the systems show similar trends across different groups. Some grammatical features impose strong challenges on the MT systems such as PathPP, ApprCLF, KindCLF, LexNeg, etc., while some other grammatical features are easier for MT systems to address, like ZAIProg, MEINeg, NPI, etc. We also see that Ernie, Google, Claude, and GPT give high performance on sentence groups of all grammatical features while Qwen performs obviously the worst among all the systems. The detailed statistics can be found in Table 15. The performance in other metrics can be found in Appendix B.

It is therefore indicated that the presence of certain grammatical features will potentially affect the performance of MT systems. We assess the impact of each grammatical feature by conducting a t-test between the MT performance on the sentence group containing the target grammatical feature and that on the remaining sentences. The result is shown in Figure 2. There are ten grammatical features imposing significant negative effects on certain MT systems: PathPP, NP, Rel, KindCLF, PtcpPP, LexNeg, PreVPP, TmpSCpl, LocPP, and Cmpr; and there are nine grammatical features having significant positive effects instead: ZAIProg, MEINeg, NPI, Recp, ZHIQtf, AdvP, Refl, SHICop, and VP. However, it is worth noting that the low scores on certain grammatical feature groups are not necessarily occasioned by the translation errors that are directly linked to the units marking the grammatical features. There are also many other implicit factors indirectly associated with the grammatical features being worthy of exploration, like the semantic or syntactic complexity.

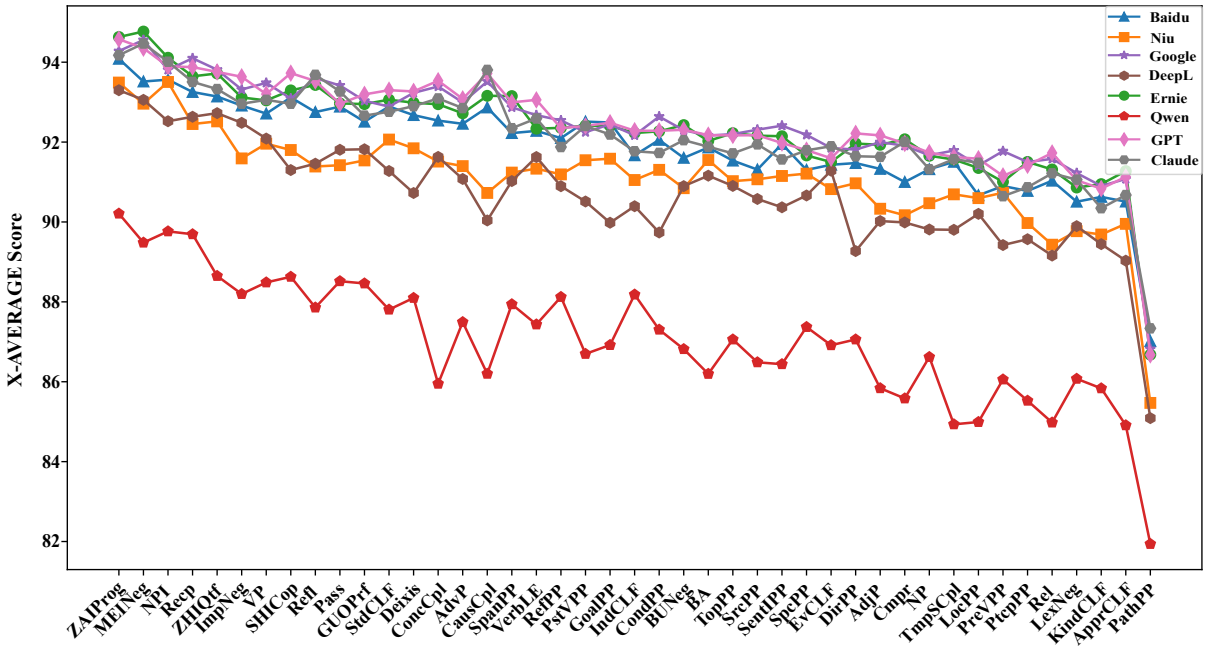


Figure 1: Performance in X-AVERAGE scores of each system on different grammatical features.

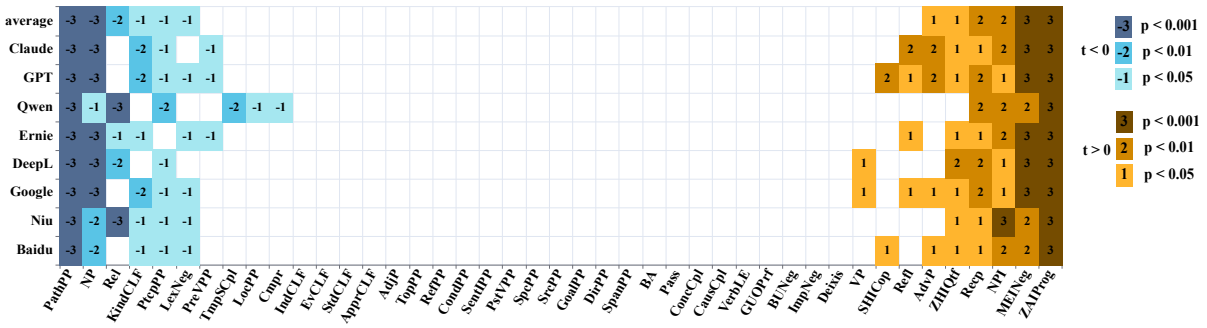


Figure 2: The significance levels of the effects of different grammatical features on the performance of each system according to t-tests in X-AVERAGE scores between the paired sentence groups with and without each grammatical feature. The blue squares on the left mean that the sentences containing a certain grammatical feature tend to get significantly lower scores than the sentences without them, and the yellow squares on the right vice versa. A significant difference with p less than 0.05, 0.01, and 0.001 are marked as 1, 2, and 3 respectively.

Since the most advanced automatic metrics are based on semantic vectors, it requires some consideration of why certain translations receive low scores, particularly in a specific sentence group. The following are some case studies to provide hints of the possible cause of the low translation scores. We should also keep in mind that this study is not aimed to provide a full understanding of why grammatical features can affect MT systems, but instead provide a different aspect and dimension for diagnosing MT systems in fine-grained levels.

4.3.2 Case Studies

In the process of translation, the accurate conveyance of meaning from the source language to the target language is important. However, errors often arise due to misinterpretation of some grammatical features, leading to mistranslation that may alter the meaning of the original text. This case study manually examines several specific instances of such errors in the translation.

One of the primary issues in automatic translation arises from the misinterpretation of path preposition phrases (PathPPs), as shown in Figure 2. Table 4 shows a typical example, where the preposition 沿 *yan* 'along' was misinterpreted, resulting in mistranslations by several MT systems. The original sentence uses the PathPP 沿 渤海 公路 *yan Bohai gong lu* 'along Bohai Highway' to describe how 唐海 *Tanghai* (a town) extends from east to west.

Source Text	Reference	Ernie	Qwen	Claude
唐海地处唐山“金三角”中央地带， 沿渤海公路贯穿东西，	Tanghai is located in the central area of Tangshan's “Golden Triangle”, running east and west along the Bohai Highway.	Tanghai is located in the central area of Tangshan's “Golden Triangle”. The Bohai Highway runs through the east and west.	Tanghai is in the heart of Tangshan's Golden Triangle, linked by the Bohai coastal highway.	Tanghai is located in the central area of the “Golden Triangle” of Tangshan. The Bohai Highway runs through the county from east to west.

Table 4: Example of mistranslation caused by grammatical feature: path preposition phrase (PathPP).

Source Text	Reference	Niu	Qwen	DeepL
如果什么东西是充足的它就是不令人羡慕的，羡慕的是缺乏的。	When something is in plenty it is not admired, but admired in case of scarcity.	If something is sufficient, it is not enviable, and enviable is lacking.	What is abundant is unenvied; it is the absence that counts.	If something is sufficient it is not enviable, envy is lacking.

Table 5: Example of mistranslation caused by the grammatical feature: relative construction (Rel).

However, in translations by Ernie and Claude, the sentence was incorrectly rendered as ‘The Bohai Highway runs through the east and west’. Qwen's translation, nevertheless, used ‘linked by’ to describe this relationship, which greatly shifted the meaning of the source sentence.

Additionally, the translation quality was significantly affected by the relative construction (Rel) (see in Figure 2). Table 5 shows a typical example of relative construction. In the source sentence, 羡慕的 *xian mu de* ‘(things) that are enviable’ is a headless clause where the elliptic head noun refers to 东西 *dong xi* ‘things’ mentioned earlier. Niu's translation misinterpreted 羡慕的 *xian mu de* as an adjectival phrase rather than a subject of the relative construction, and thus misunderstood the meaning of the original sentence. Qwen just omitted the real subject -羡慕的(东西)- of the sub-clause, leading to mistranslating the adjective 缺乏的 *que fa de* ‘scarce’ as the subject. DeepL's translation completely ignored the relative construction marker 的 ‘*de*’ and treated 羡慕 *xian mu* ‘envy’ as the subject.

Another grammatical point that has a significant negative impact is noun phrases (NPs). NP has a relatively large number of sentences on Laws and Thesis (see in Table 1), which have generally low averaged performance (see in Table 3) possibly due to their high semantic complexity regarding professionalism. This partially explains the negative effects of NP. Particularly, the specialized terminology within the Thesis category can notably contribute to the translation challenges.

5 Additional Discussion

In this section, we discuss several potential interfering factors that may also affect the quality of automatic translation by interacting with grammatical features, including sentence length, domain, and the effects of different automatic metrics.

5.1 Analysis of Sentence Length

It has long been an observed consensus that longer sentences are generally more difficult to MT systems and thus result in lower qualities and scores (Cho et al., 2014; Koehn and Knowles, 2017). This can also be verified by the significant inverse relationship between the lengths of source sentences and their translation scores given by human experts as shown in Figure 3, generated on WMT23 data (Freitag et al., 2023).

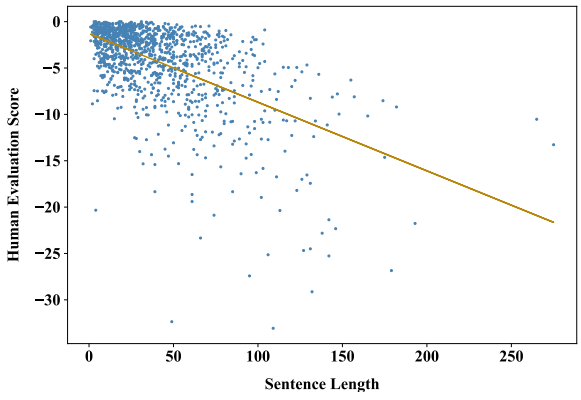


Figure 3: The correlation between sentence lengths (x-axis) and human average translation scores (y-axis) on WMT23 Chinese-English dataset for the metric shared task.

To assess whether the significant effects of grammatical features as observed are due to differences in sentence lengths among the groups, we calculated the average sentence length for each sentence group with a specific grammatical feature and examined the relationship between average sentence lengths and the corresponding X-AVERAGE scores. From the result, as shown in Figure 4, we can see that although certain sentence groups of different grammatical features have different sentence lengths ranging from 19.39 to 32.18 characters, they do not show significant correlation with the average X-AVERAGE scores of the groups, indicating that effects by grammatical features are not due to the bias of sentence length distribution.

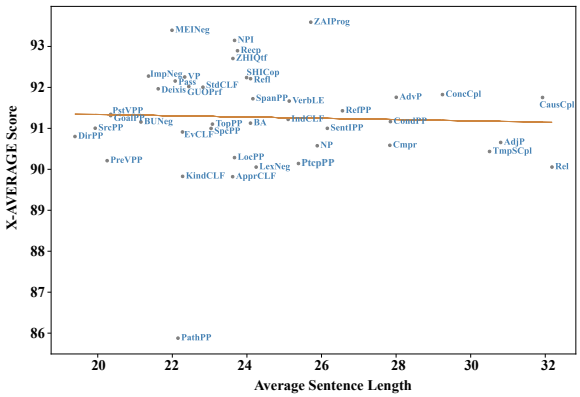


Figure 4: The correlation between average sentence lengths and X-AVERAGE scores of sentence groups containing different grammatical features with Pearson'r = -0.039 and p = 0.802.

5.2 Analysis of Domains

As mentioned in Section 4.3, domains have a significant influence on the performance of MT systems due to variations in vocabulary and writing registers. Therefore, it is important to consider whether domains have contributed to the observed significant impact of certain grammatical features on MT systems. Reviewing the data in Table 1, we make all grammatical feature groups maintain a roughly balanced distribution in terms of sentence number across the seven domains except Rel, which extensively exists in the domain of Laws. This balance allows us to focus more on the effects of grammatical features rather than domains when calculating statistics between feature-accordingly grouped sentences.

Interestingly, the similar trends of MT systems' performance across different domains may also correlate with other factors e.g. sentence length. Thus, a further question is whether the effects apparently imposed by domains on the scores of MT systems are partially due to the imbalanced distribution of sentence length across domains. Table 6 displays the average sentence lengths of different domains along with their standard deviations. We see that Spoken and Subtitles have the shortest sentences while Laws and Thesis have the longest ones, with a gap of about 20 characters between them. This may partially explain why MT systems achieve the best performance when rendering materials in the former two domains while the worst is in the latter two domains as shown in Table 3.

Edu	Laws	News	Sci	Spk	Sbt	Ths
23±7	30±11	25±10	21±10	18±6	15±4	29±12

Table 6: Average sentence length of each domain with the standard deviation.

5.3 Analysis of Evaluation Metrics

Following the hypothesis of regarding human evaluation as the gold standard, the metrics that generate judgments on translation quality more similar to humans are superior (Freitag et al., 2023).

In Figure 3, we see that there exists a significant negative correlation relationship between sentence lengths and human evaluation scores. To know if different metrics rate MT qualities similarly regarding sentence length, we examine the correlations between sentence lengths and scores generated by six metrics to meta-evaluate their effectiveness. We provide scatter plots of sentence lengths and scores based on both WMT23 data (Freitag et al., 2023) and our data, and the results are presented in Figure 5. On both datasets, XCOMET and XCOMET-QE exhibit patterns similar to human evaluations and are therefore considered to provide more reliable scores, particularly regarding the negative effects of sentence length. However, BLEU, CHRF, COMET, and COMET-QE yield judgments on translation quality that are inconsistent with human evaluations. According to Table 7, CHRF, COMET, and COMET-QE even exhibit significant positive correlations, indicating their bias towards longer sentences. This finding is consistent with the leaderboard of metrics concluded by WMT23 (Freitag et al., 2023), which ranks XCOMET and XCOMET-QE as the top performers.

Besides, Table 7 shows that the average system

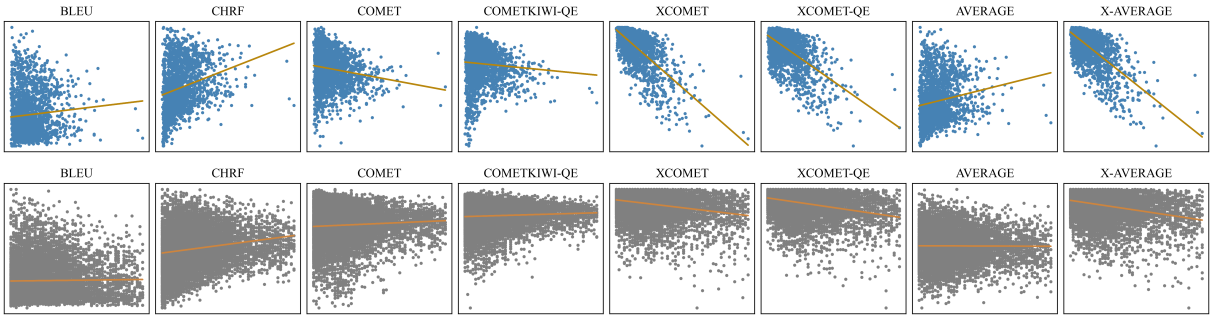


Figure 5: The correlation between sentence lengths (x-axis) and average translation scores (y-axis) in different measures. XCOMET and XCOMET-QE are more consistent with the human evaluation trend in Figure 3. The first row is on the WMT23 Chinese-English dataset for the metric shared task and the second row is on our test suite. Metrics show similar patterns on the two different datasets except for COMET and COMETKIWI-QE.

Metric	Pearson'r	P-value
BLEU	0.016	0.149
CHRF	0.185	*** 0.000
COMET	0.067	*** 0.000
COMETKIWI-QE	0.064	*** 0.000
XCOMET	-0.242	*** 0.000
XCOMET-QE	-0.302	*** 0.000
AVERAGE	-0.006	0.581
X-AVERAGE	-0.286	*** 0.000

Table 7: The correlation between sentence lengths and system average scores in different metrics on our test suite.

scores of all six automatic metrics (AVERAGE) do not significantly correlate with sentence length by offsetting the effects of different metrics. On the contrary, the average scores of XCOMET and XCOMET-QE (X-AVERAGE) remain the high reliability by showing a significant negative correlation between scores and sentence lengths. Therefore, XCOMET, XCOMET-QE, and X-AVERAGE are more recommended for practical evaluation.

6 Conclusion and Future Work

In this paper, we investigate the impact of various grammatical features (linguistic phenomena) on eight state-of-the-art NMT systems and LLMs with a test suite we newly constructed. Although LLMs have achieved promising performance on many NLP tasks, NMT systems especially Google Translate have outperformed most of the LLMs in the Chinese-English automatic translation task. It is observed that certain grammatical features pose a great challenge to NMT systems and LLMs including the ones developed by Chinese companies such as Baidu, Ernie, Niu, and Qwen. We also discuss other possible factors that may also impact MT systems including sentence length, domain, and the evaluation metrics. We find that the Thesis category is particularly more difficult due to its comparatively longer sentence and the existence of a large number of terminologies. In addition, we confirm that longer sentences are generally more difficult for MT systems. However, our analysis of the correlation between the sentence length and different metrics reveals that BLEU and CHRF tend to rate shorter sentences with lower scores, which is contradictory to human evaluation. This also confirms that XCOMET and XCOMET-QE are the most reliable metrics according to the results of the WMT23 metrics shared task.

Currently, our test suite does not cover all the 157 grammatical features of Chinese due to the rareness of some particular grammatical features. In the future, we plan to extend our test suite to cover all the grammatical features by resorting to other resources.

Limitations

One limitation of our study is the absence of human evaluation scores. Our analysis relies heavily on automatic metrics, specifically the average score of XCOMET and XCOMET-QE. The former relies on reference translations and the latter does not. According to the WMT23 metrics shared task results (Freitag et al., 2023), both metrics show a very high correlation with human scores. This demonstrates the validity and reliability of data in our study to some extent. While human evaluation is the most reliable, it is also expensive and impractical for assessing every MT system. In contrast, the test suite, combined with automatic evaluation metrics, offers a convenient and efficient tool for evaluating any MT systems, providing immediate

diagnostic reports.

Another limitation of our study is that it does not cover all the grammatical features of Chinese due to the scarcity of certain grammatical features. We plan to address this issue by exploring other data sources to cover all other grammatical features in the future.

Acknowledgements

Firstly, we thank all reviewers for their valuable and insightful comments. We also acknowledge the funding support from the Supervisor Academic Guidance Program of Shanghai International Studies University (Grant No. 2022113024) and the Linguistics Frontier Research Funding of Shanghai International Studies University (Grant No. 41004525/001).

References

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Lorna Balkan. 1994. Test suites: some issues in their use and design. In *Proceedings of the Second International Conference on Machine Translation: Ten years on*, Cranfield University, UK.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Peter Jan-Thorsten, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural mt engines. *The Prague bulletin of mathematical linguistics*, 108(1):159.

Xinyi Cai and Deyi Xiong. 2020. A test suite for evaluating discourse phenomena in document-level neural machine translation. In *Proceedings of the Second International Workshop of Discourse Processing*, pages 13–17, Suzhou, China. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Multifaceted challenge set for evaluating machine translation performance. In *Proceedings of the Eighth Conference on Machine Translation*, pages 217–223, Singapore. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Nuno Guerreiro, Ricardo Rei, Daan Van, Pierre Colombo, Luisa Coheur, and André Martins. 2023. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Chu-Ren Huang and Dingxu Shi, editors. 2016. *A Reference Grammar of Chinese*. Cambridge University Press, Cambridge.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Qianchu Liu, Federico Fancellu, and Bonnie Webber. 2018. NegPar: A parallel corpus annotated for negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Fine-grained evaluation of German-English machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073, Online. Association for Computational Linguistics.

Vivien Macketanz, Shushen Manakhimova, Eleftherios Avramidis, Ekaterina Lapshinova-koltunski, Sergei Bagdasarov, and Sebastian Möller. 2022. Linguistically motivated evaluation of the 2022 state-of-the-art machine translation systems for three language directions. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 432–449, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2023. IIIT HYD's submission for WMT23 test-suite task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 246–251, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2019. Evaluating conjunction disambiguation on English-to-German and French-to-German WMT 2019 translation hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 464–469, Florence, Italy. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. 2019. A test suite and manual evaluation of document-level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy. Association for Computational Linguistics.

Huacheng Song and Hongzhi Xu. 2024a. Benchmarking the performance of machine translation evaluation metrics with Chinese multiword expressions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2204–2216, Torino, Italia. ELRA and ICCL.

Huacheng Song and Hongzhi Xu. 2024b. A deep analysis of the impact of multiword expressions and named entities on Chinese-English machine translations. In

Proceedings of the 2024 Conference on Emprical Methods on Natural Language Processing (EMNLP 2024).

Gongbo Tang, Philipp Rönchen, Rico Sennrich, and Joakim Nivre. 2021. Revisiting negation in neural machine translation. *Transactions of the Association for Computational Linguistics*, 9:740–755.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. UM-corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hongzhi Xu and Jingxia Lin. 2023. A reference grammar based chinese corpus. In *Symposium on Language and Big Data: Challenges in Chinese Linguistics*. Hong Kong.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark.

A System Performance on Different Domains

Table 8 to Table 14 show the performance of different systems on different domains in different metrics. In all the following tables, the highest and the lowest scores among all systems evaluated are highlighted with bold and underlined numbers.

	Edu	Laws	News	Sci	Spk	Sbt	Ths
Baidu	69.1	71.0	70.5	70.6	69.3	67.2	66.0
Niu	69.3	73.2	70.9	71.2	69.5	67.8	66.0
Google	**71.0**	**75.3**	**72.6**	**72.4**	**70.5**	68.6	**67.0**
DeepL	69.0	71.9	70.9	70.2	69.1	**68.8**	65.0
Ernie	70.1	72.8	71.4	71.7	69.9	68.3	**67.0**
Qwen	62.7	64.5	64.4	64.6	63.6	63.7	59.5
GPT	70.2	71.0	71.4	71.5	69.7	68.4	66.4
Claude	70.2	72.3	71.5	71.4	69.6	68.2	66.0
NMT-AVG	69.6	72.8	71.2	71.1	69.6	68.1	66.0
LLM-AVG	68.3	70.2	69.7	69.8	68.2	67.2	64.7

Table 8: Performance in AVERAGE scores of each system on different domains.

	Edu	Laws	News	Sci	Spk	Sbt	Ths
Baidu	20.4	25.8	22.2	24.8	20.6	19.5	17.8
Niu	22.8	35.8	24.6	28.0	22.7	22.0	19.5
Google	**24.8**	**40.2**	**28.0**	**29.7**	**23.6**	22.9	19.6
DeepL	22.4	32.9	24.9	25.9	22.2	**25.6**	16.6
Ernie	22.6	31.0	23.9	26.9	21.7	21.9	**19.9**
Qwen	15.4	18.8	17.0	18.3	16.0	19.0	11.8
GPT	22.2	25.2	23.4	26.1	21.0	22.0	17.8
Claude	22.4	30.6	23.9	26.6	20.8	21.7	17.2
NMT-AVG	22.6	33.7	24.9	27.1	22.3	22.5	18.4
LLM-AVG	20.6	26.4	22.0	24.5	19.9	21.1	16.7

Table 9: Performance in BLEU scores of each system on different domains.

	Edu	Laws	News	Sci	Spk	Sbt	Ths
Baidu	55.1	59.2	56.6	59.4	52.8	48.1	59.0
Niu	57.0	66.3	58.1	61.5	54.3	50.5	60.4
Google	**58.5**	**69.1**	**60.5**	**62.5**	**55.1**	50.6	60.5
DeepL	56.3	64.9	58.1	59.4	53.8	**52.5**	57.3
Ernie	57.0	63.4	58.1	61.4	53.9	50.0	**61.2**
Qwen	45.5	49.7	47.2	49.6	44.3	44.3	47.8
GPT	57.0	58.8	57.9	61.0	53.1	50.1	59.6
Claude	57.2	63.4	58.4	61.3	53.3	49.8	59.2
NMT-AVG	56.7	64.9	58.3	60.7	54.0	50.4	59.3
LLM-AVG	54.2	58.8	55.4	58.3	51.1	48.5	57.0

Table 10: Performance in CHRF scores of each system on different domains.

B System Performance on Different Grammatical Features

Table 15 to Table 22 show the performance of different systems on different grammatical feature groups in different metrics.

	Edu	Laws	News	Sci	Spk	Sbt	Ths
Baidu	81.3	83.3	83.0	82.0	81.3	78.2	76.1
Niu	81.2	84.2	83.1	82.2	81.3	78.3	76.0
Google	**82.8**	**86.0**	**84.5**	**83.0**	**82.2**	**79.2**	76.5
DeepL	81.0	83.2	83.1	81.1	80.8	78.8	74.9
Ernie	82.3	84.4	83.8	82.5	81.7	78.8	**76.6**
Qwen	76.4	79.1	78.7	77.2	76.3	74.9	71.7
GPT	82.4	83.3	83.9	82.4	81.6	78.8	76.5
Claude	82.2	84.0	83.9	82.4	81.5	78.6	76.4
NMT-AVG	81.6	84.2	83.4	82.1	81.4	78.6	75.9
LLM-AVG	80.8	82.7	82.6	81.1	80.3	77.8	75.3

Table 11: Performance in COMET scores of each system on different domains.

	Edu	Laws	News	Sci	Spk	Sbt	Ths
Baidu	90.2	90.9	91.9	91.3	92.5	90.9	84.7
Niu	89.7	90.4	91.5	91.2	92.1	90.6	83.8
Google	**91.9**	**92.5**	**93.2**	**92.6**	**93.4**	**91.8**	85.9
DeepL	89.3	89.1	91.3	90.5	91.7	91.3	83.4
Ernie	91.4	92.0	92.9	92.4	93.2	**91.8**	**86.0**
Qwen	84.9	85.4	86.8	87.0	88.1	87.6	78.6
GPT	91.6	91.3	93.1	92.3	93.1	**91.8**	85.3
Claude	91.5	91.2	92.9	91.9	92.9	91.6	85.1
NMT-AVG	90.3	90.7	92.0	91.4	92.4	91.2	84.5
LLM-AVG	89.8	90.0	91.4	90.9	91.8	90.7	83.8

Table 12: Performance in XCOMET scores of each system on different domains.

	Edu	Laws	News	Sci	Spk	Sbt	Ths
Baidu	73.8	**74.3**	74.7	72.4	72.8	70.7	70.1
Niu	72.7	71.6	74.0	71.5	72.0	70.1	68.9
Google	74.4	72.7	75.2	72.7	72.8	71.5	70.4
DeepL	72.6	71.4	74.2	71.8	71.6	70.1	69.7
Ernie	74.1	73.7	75.3	73.0	73.2	71.9	70.1
Qwen	65.2	65.8	66.9	65.1	64.8	64.1	63.1
GPT	**74.5**	**74.3**	**75.5**	**73.3**	73.5	**72.2**	**70.7**
Claude	74.4	72.7	75.3	72.8	**73.6**	72.0	70.3
NMT-AVG	73.4	72.5	74.5	72.1	72.3	70.6	69.8
LLM-AVG	72.1	71.6	73.2	71.0	71.3	70.0	68.5

Table 13: Performance in COMETKIWI-QE scores of each system on different domains.

	Edu	Laws	News	Sci	Spk	Sbt	Ths
Baidu	93.5	92.9	94.4	93.7	95.6	95.4	88.4
Niu	92.4	90.8	93.7	93.0	94.9	94.9	87.0
Google	**93.7**	91.3	94.5	93.9	95.6	**95.6**	**88.7**
DeepL	92.2	89.6	93.7	92.7	94.5	94.5	87.7
Ernie	93.4	92.4	94.6	93.8	95.6	95.4	88.2
Qwen	88.4	88.1	89.9	90.0	91.9	92.1	84.1
GPT	**93.7**	**93.3**	**94.8**	**94.0**	**95.8**	**95.6**	88.4
Claude	93.6	92.3	94.5	93.6	95.5	95.2	87.9
NMT-AVG	93.0	91.2	94.1	93.3	95.2	95.1	88.0
LLM-AVG	92.3	91.5	93.5	92.8	94.7	94.6	87.2

Table 14: Performance in XCOMET-QE scores of each system on different domains.

	Baidu	Niu	Google	DeepL	Ernie	Qwen	GPT	Claude	NMT-AVG	LLM-AVG	All-AVG
ZAIProg	94.1	93.5	94.3	93.3	**94.6**	90.2	**94.6**	94.2	93.8	93.4	93.6
MEINeg	93.5	93.0	94.6	93.1	**94.8**	89.5	94.4	94.5	93.6	93.3	93.4
NPI	93.6	93.5	93.8	92.5	**94.1**	89.8	93.9	94.0	93.3	92.9	93.1
Recp	93.3	92.5	**94.1**	92.6	93.6	89.7	93.9	93.5	93.1	92.7	92.9
ZHIQtf	93.1	92.5	**93.8**	92.7	93.7	88.6	**93.8**	93.3	93.0	92.4	92.7
ImpNeg	92.9	91.6	93.3	92.5	93.1	88.2	**93.6**	93.0	92.6	92.0	92.3
VP	92.7	92.0	**93.5**	92.1	93.0	88.5	93.2	93.1	92.6	91.9	92.3
SHICop	93.1	91.8	93.1	91.3	93.3	88.6	**93.7**	93.0	92.3	92.1	92.2
Refl	92.8	91.4	93.6	91.5	93.4	87.9	93.5	**93.7**	92.3	92.1	92.2
Pass	92.9	91.4	**93.4**	91.8	93.0	88.5	93.0	93.3	92.4	92.0	92.2
GUOPrf	92.5	91.5	93.0	91.8	93.0	88.5	**93.2**	92.7	92.2	91.8	92.0
StdCLF	92.9	92.1	92.9	91.3	93.1	87.8	**93.3**	92.8	92.3	91.8	92.0
Deixis	92.7	91.8	93.2	90.7	93.0	88.1	**93.3**	92.9	92.1	91.8	92.0
ConcCpl	92.5	91.5	93.4	91.6	92.9	85.9	**93.5**	93.1	92.2	91.3	91.8
AdvP	92.5	91.4	93.0	91.1	92.7	87.5	**93.1**	92.9	92.0	91.5	91.8
CausCpl	92.9	90.7	93.5	90.0	93.2	86.2	93.7	**93.8**	91.8	91.7	91.8
SpanPP	92.2	91.2	92.9	91.0	**93.2**	87.9	93.0	92.3	91.8	91.6	91.7
VerbLE	92.3	91.3	92.7	91.6	92.3	87.4	**93.1**	92.6	92.0	91.3	91.7
RefPP	92.1	91.2	**92.5**	90.9	92.4	88.1	92.3	91.9	91.7	91.2	91.4
PstVPP	**92.5**	91.5	92.2	90.5	92.4	86.7	92.4	92.4	91.7	91.0	91.3
GoalPP	**92.5**	91.6	92.4	90.0	92.4	86.9	**92.5**	92.2	91.6	91.0	91.3
IndCLF	91.7	91.1	92.2	90.4	92.2	88.2	**92.3**	91.8	91.3	91.1	91.2
CondPP	92.1	91.3	**92.6**	89.7	92.3	87.3	92.3	91.7	91.4	90.9	91.2
BUNeg	91.6	90.8	92.3	90.9	**92.4**	86.8	92.3	92.0	91.4	90.9	91.2
BA	91.9	91.6	**92.2**	91.2	92.0	86.2	**92.2**	91.9	91.7	90.6	91.1
TopPP	91.5	91.0	**92.2**	90.9	**92.2**	87.1	**92.2**	91.7	91.4	90.8	91.1
SrcPP	91.3	91.1	**92.3**	90.6	92.2	86.5	92.2	91.9	91.3	90.7	91.0
SentIPP	92.0	91.2	**92.4**	90.4	92.1	86.4	92.0	91.6	91.5	90.5	91.0
SpcPP	91.3	91.2	**92.2**	90.7	91.7	87.4	91.8	91.8	91.3	90.7	91.0
EvCLF	91.4	90.8	**91.9**	91.3	91.5	86.9	91.6	**91.9**	91.4	90.5	90.9
DirPP	91.5	91.0	91.8	89.3	92.0	87.1	**92.2**	91.6	90.9	90.7	90.8
AdjP	91.3	90.3	92.0	90.0	91.9	85.8	**92.2**	91.6	90.9	90.4	90.7
Cmpr	91.0	90.2	91.9	90.0	**92.1**	85.6	92.0	92.0	90.8	90.4	90.6
NP	91.3	90.5	**91.7**	89.8	**91.7**	86.6	**91.7**	91.3	90.8	90.3	90.6
TmpSCpl	91.5	90.7	**91.8**	89.8	91.6	84.9	91.6	91.6	91.0	89.9	90.4
LocPP	90.7	90.6	91.4	90.2	91.4	85.0	**91.6**	91.5	90.7	89.9	90.3
PreVPP	90.9	90.7	**91.8**	89.4	91.0	86.1	91.2	90.6	90.7	89.7	90.2
AgtPP	90.8	90.0	**91.5**	89.6	**91.5**	85.5	91.4	90.9	90.5	89.8	90.1
Rel	91.0	89.4	91.6	89.2	91.3	85.0	**91.7**	91.2	90.3	89.8	90.1
LexNeg	90.5	89.8	**91.2**	89.9	90.9	86.1	91.0	91.1	90.3	89.8	90.1
KindCLF	90.6	89.7	**90.9**	89.4	**90.9**	85.8	90.8	90.3	90.2	89.5	89.8
ApprCLF	90.5	90.0	91.1	89.0	**91.3**	84.9	91.1	90.7	90.2	89.5	89.8
PathPP	87.0	85.5	86.8	85.1	86.7	81.9	86.7	**87.3**	86.1	85.7	85.9

Table 15: Performance in X-AVERAGE scores of each system on different linguistic features.

	Baidu	Niu	Google	DeepL	Ernie	Qwen	GPT	Claude	NMT-AVG	LLM-AVG	All-AVG
CondPP	70.9	72.7	**74.1**	70.8	72.1	64.1	70.9	71.5	72.1	69.7	70.9
SpanPP	70.7	71.2	**72.7**	70.7	72.3	65.0	71.7	71.6	71.3	70.2	70.7
ZAIProg	70.6	71.3	**72.9**	71.0	71.8	65.4	71.3	71.1	71.4	69.9	70.7
SrcPP	70.6	71.2	**72.5**	69.9	71.7	65.0	71.4	71.7	71.1	70.0	70.5
SpcPP	69.8	71.2	**72.7**	71.0	71.2	65.0	70.8	70.9	71.2	69.5	70.3
LocPP	70.0	72.1	**73.0**	70.7	71.3	63.5	70.7	71.2	71.5	69.2	70.3
ImpNeg	70.3	70.2	**73.3**	70.8	71.1	63.9	70.9	71.0	71.2	69.2	70.2
GUOPrf	70.0	70.4	**71.6**	70.1	71.1	65.3	71.2	70.9	70.5	69.6	70.1
NPI	69.9	71.2	**72.3**	70.7	70.9	63.7	70.3	71.1	71.0	69.0	70.0
IndCLF	69.8	70.4	**72.2**	70.3	70.9	65.0	70.7	70.7	70.7	69.3	70.0
MEINeg	69.5	70.4	**72.2**	70.5	71.4	63.7	70.8	70.9	70.7	69.2	69.9
TmpSCpl	70.2	70.9	**71.7**	70.3	71.0	63.8	70.5	71.0	70.8	69.1	69.9
Recp	69.9	70.4	**72.0**	70.3	70.8	64.1	70.6	71.0	70.7	69.1	69.9
StdCLF	69.8	71.1	**71.9**	70.0	71.7	64.1	70.2	70.1	70.7	69.0	69.9
Pass	70.0	70.6	**72.0**	70.3	70.5	64.3	69.8	70.7	70.7	68.8	69.8
BA	69.6	71.0	**72.1**	70.4	70.6	63.0	70.5	70.2	70.8	68.6	69.7
RefPP	69.9	70.3	**71.4**	69.6	70.4	64.3	70.0	70.2	70.3	68.7	69.5
EvCLF	69.7	69.8	**71.3**	70.5	70.4	63.5	69.7	70.4	70.3	68.5	69.4
PstVPP	69.4	70.0	**70.9**	69.6	70.2	63.4	70.2	70.7	70.0	68.6	69.3
LexNeg	68.9	69.5	**71.3**	69.6	70.7	64.0	69.9	70.4	69.8	68.8	69.3
ZHIQtf	69.3	69.7	**71.1**	70.3	70.1	63.4	70.3	70.0	70.1	68.5	69.3
Rel	69.3	70.1	**72.1**	69.4	70.3	62.6	69.4	70.5	70.2	68.2	69.2
NP	69.2	70.0	**71.5**	69.4	70.3	63.4	69.7	70.0	70.0	68.3	69.2
VP	69.1	69.9	**71.5**	69.4	70.1	63.6	69.8	70.1	70.0	68.4	69.2
AdvP	69.1	69.6	**71.3**	69.3	70.3	62.8	69.9	70.2	69.8	68.3	69.1
Deixis	69.1	69.7	**71.3**	69.0	70.3	63.6	69.7	69.9	69.8	68.4	69.1
GoalPP	69.0	69.5	**70.7**	69.1	69.9	63.5	69.9	70.3	69.6	68.4	69.0
VerbLE	69.1	69.6	**70.6**	69.6	70.0	63.2	69.9	69.7	69.7	68.2	69.0
CausCpl	69.1	68.8	**71.5**	68.3	70.7	61.8	70.3	70.8	69.4	68.4	68.9
TopPP	68.7	69.5	**70.9**	69.9	70.0	62.2	69.7	69.7	69.8	67.9	68.8
SHICop	68.9	69.3	**70.8**	68.8	70.0	63.3	69.8	69.3	69.5	68.1	68.8
DirPP	68.3	69.8	**70.4**	67.4	69.3	62.8	69.2	69.4	69.0	67.7	68.3
PtcpPP	68.4	69.1	**70.6**	68.5	69.2	62.5	68.7	68.9	69.2	67.3	68.3
Refl	68.2	68.2	**70.1**	68.3	69.3	62.4	69.1	69.7	68.7	67.6	68.2
Cmpr	67.9	68.4	**70.3**	68.5	69.6	61.3	68.7	69.0	68.8	67.1	68.0
PreVPP	68.2	68.9	**70.3**	67.7	69.0	61.9	68.8	68.9	68.8	67.2	68.0
BUNeg	67.8	68.8	**70.2**	68.3	69.0	61.7	68.4	68.3	68.8	66.8	67.8
KindCLF	67.5	68.4	**69.6**	67.8	68.7	62.1	68.3	68.4	68.3	66.9	67.6
AdjP	67.7	68.4	**69.7**	68.0	68.8	61.1	68.3	68.5	68.5	66.7	67.6
SentIPP	68.0	68.2	**69.0**	67.4	68.1	60.8	67.8	67.7	68.2	66.1	67.1
ConcCpl	67.2	67.3	**69.0**	67.1	68.1	60.1	68.5	68.1	67.7	66.2	66.9
ApprCLF	67.2	67.4	**69.2**	66.9	67.6	60.1	67.6	68.1	67.7	65.8	66.7
PathPP	66.4	66.8	**67.9**	66.0	67.6	60.1	66.1	67.1	66.8	65.2	66.0

Table 16: Performance in AVERAGE scores of each system on different grammatical features.

	Baidu	Niu	Google	DeepL	Ernie	Qwen	GPT	Claude	NMT-AVG	LLM-AVG	All-AVG
CondPP	25.1	33.4	**36.2**	29.7	28.6	17.2	25.1	28.0	31.1	24.7	27.9
LocPP	24.4	31.8	**33.1**	28.3	27.5	18.4	25.7	27.6	29.4	24.8	27.1
SrcPP	25.6	28.7	**30.9**	26.4	27.8	21.2	26.2	27.4	27.9	25.6	26.8
IndCLF	24.2	27.5	**31.0**	28.5	26.7	19.6	25.7	26.3	27.8	24.6	26.2
SpcPP	23.4	28.3	**30.8**	28.6	26.9	19.1	25.3	26.0	27.8	24.3	26.0
SpanPP	24.5	27.7	**30.0**	26.5	28.1	18.4	26.3	26.6	27.2	24.9	26.0
TmpSCpl	24.2	27.9	**29.1**	27.4	26.2	18.9	24.4	26.2	27.1	23.9	25.5
ImpNeg	23.4	25.2	**32.6**	26.2	25.9	17.5	24.6	26.3	26.8	23.6	25.2
GUOPrf	23.2	26.4	**27.9**	26.2	25.3	20.8	25.6	25.4	25.9	24.3	25.1
EvCLF	24.8	26.1	**28.9**	27.1	26.4	18.3	23.8	25.1	26.7	23.4	25.1
PstVPP	22.2	26.1	**27.4**	26.8	24.6	18.9	24.6	25.9	25.6	23.5	24.6
NP	22.4	26.7	**29.5**	26.1	25.3	17.6	23.1	24.6	26.2	22.6	24.4
Pass	23.0	28.0	**28.2**	26.1	23.6	19.3	21.5	23.9	26.3	22.1	24.2
Rel	22.3	27.7	**30.6**	25.4	24.3	16.4	21.1	25.2	26.5	21.8	24.1
TopPP	21.6	25.5	**28.1**	27.4	25.2	15.7	23.9	24.7	25.6	22.4	24.0
NPI	21.7	26.6	**29.4**	26.8	24.1	15.1	22.3	24.8	26.1	21.6	23.9
StdCLF	21.1	28.0	**28.3**	25.1	26.0	17.7	21.8	22.8	25.6	22.1	23.8
LexNeg	21.1	25.2	**28.5**	24.8	25.5	17.5	23.1	24.8	24.9	22.7	23.8
Deixis	22.1	25.8	**28.7**	25.2	25.0	17.2	22.4	23.9	25.5	22.1	23.8
GoalPP	21.2	24.5	**26.1**	25.9	23.4	19.3	23.5	25.0	24.4	22.8	23.6
ZAIProg	22.1	25.4	**28.9**	25.1	24.3	16.8	23.1	23.1	25.4	21.8	23.6
Recp	22.3	25.4	**27.3**	24.4	24.1	16.6	22.7	24.7	24.9	22.0	23.4
MEINeg	20.7	25.1	**27.3**	24.7	25.0	16.4	23.0	23.4	24.4	22.0	23.2
BA	21.0	25.7	**28.5**	25.4	23.6	15.7	23.2	22.6	25.1	21.3	23.2
DirPP	20.8	**27.0**	**27.0**	21.8	22.0	17.7	21.8	23.2	24.1	21.2	22.7
KindCLF	20.5	25.2	**26.9**	23.7	23.4	16.6	21.7	23.1	24.1	21.2	22.6
ZHIQtf	21.3	24.0	**25.7**	25.4	21.9	17.0	22.4	22.2	24.1	20.9	22.5
RefPP	21.7	24.9	**26.3**	23.1	23.0	16.0	21.4	22.4	24.0	20.7	22.4
VP	20.8	24.8	**26.9**	23.0	22.4	15.5	21.3	22.3	23.9	20.4	22.1
AdvP	20.6	23.9	**26.5**	23.7	23.2	14.6	21.5	22.7	23.7	20.5	22.1
PreVPP	21.0	23.6	**25.5**	21.7	23.1	15.1	22.1	23.1	22.9	20.9	21.9
PathPP	20.6	23.9	**25.8**	22.4	24.6	15.0	19.5	21.4	23.2	20.1	21.7
SHICop	19.6	23.3	**25.4**	22.9	22.6	15.5	20.9	21.1	22.8	20.0	21.4
ApprCLF	20.3	21.7	**26.4**	22.2	20.5	14.5	20.9	22.6	22.7	19.6	21.1
BUNeg	19.8	24.2	**25.9**	23.1	21.6	14.7	19.6	20.1	23.2	19.0	21.1
Cmpr	19.5	22.6	**25.7**	23.0	22.8	14.5	19.9	20.4	22.7	19.4	21.0
VerbLE	19.7	23.3	**23.6**	22.3	22.1	15.3	20.5	20.5	22.2	19.6	20.9
PtcpPP	19.7	23.3	**25.7**	22.4	21.1	14.3	19.2	20.8	22.8	18.9	20.8
CausCpl	19.1	21.3	**25.3**	20.3	23.8	12.4	21.3	22.5	21.5	20.0	20.8
Refl	18.7	20.5	**22.4**	21.6	21.0	14.4	19.8	21.5	20.8	19.2	20.0
AdjP	17.1	20.7	**22.5**	20.1	19.3	12.6	17.1	18.4	20.1	16.9	18.5
SentIPP	17.1	**19.4**	19.2	17.9	17.5	11.5	16.1	16.9	18.4	15.5	16.9
ConcCpl	14.5	17.0	**18.4**	16.6	16.0	10.1	16.6	15.9	16.6	14.7	15.6

Table 17: Performance in BLEU scores of each system on different grammatical features.

	Baidu	Niu	Google	DeepL	Ernie	Qwen	GPT	Claude	NMT-AVG	LLM-AVG	All-AVG
LocPP	60.6	66.1	**66.7**	62.5	63.0	51.9	61.0	62.8	64.0	59.7	61.8
CondPP	60.8	65.6	**67.1**	63.6	63.4	49.7	60.9	62.7	64.3	59.2	61.7
SrcPP	60.2	62.6	**63.8**	59.8	62.7	52.8	61.8	63.0	61.6	60.1	60.9
LexNeg	59.5	61.6	**63.6**	61.0	63.1	51.8	60.9	62.0	61.4	59.5	60.4
SpcPP	57.8	61.7	**63.5**	61.3	60.8	51.5	59.8	61.0	61.1	58.3	59.7
IndCLF	58.1	60.1	**62.5**	60.3	60.0	50.6	58.8	59.7	60.2	57.3	58.8
SpanPP	57.7	60.3	**61.4**	59.4	60.7	48.7	59.1	59.7	59.7	57.0	58.4
TmpSCpl	57.1	59.9	**60.3**	60.0	59.4	48.9	57.8	59.2	59.3	56.3	57.8
RefPP	57.9	60.3	**61.0**	59.0	59.0	48.2	57.7	59.1	59.5	56.0	57.8
Rel	56.8	60.8	**62.6**	59.9	59.1	47.2	56.2	59.6	60.0	55.5	57.8
Recp	57.3	59.7	**61.2**	59.1	59.0	47.7	58.1	59.5	59.3	56.1	57.7
BA	56.8	60.5	**62.1**	59.3	58.9	47.0	57.8	58.2	59.7	55.5	57.6
EvCLF	57.2	58.5	**60.3**	59.8	59.2	48.6	57.5	59.0	59.0	56.1	57.5
TopPP	56.7	59.7	**61.1**	60.2	58.8	46.5	58.1	58.9	59.4	55.6	57.5
NP	56.6	59.7	**61.5**	59.0	59.1	48.2	57.2	58.6	59.2	55.8	57.5
GUOPrf	56.7	58.9	**59.6**	58.3	59.0	49.6	58.7	58.8	58.4	56.5	57.4
ImpNeg	56.5	59.3	**62.7**	58.4	58.3	47.0	56.8	58.5	59.2	55.1	57.2
StdCLF	55.8	60.0	**60.2**	57.7	60.1	47.9	56.1	56.4	58.4	55.1	56.8
ZAIProg	56.2	58.4	**60.8**	57.7	58.2	47.2	57.1	57.2	58.3	54.9	56.6
Pass	55.7	59.4	**59.9**	57.8	57.8	47.8	56.0	57.6	58.2	54.8	56.5
PstVPP	55.2	58.3	**58.8**	**58.8**	57.0	47.5	57.2	58.5	57.8	55.0	56.4
PtcpPP	55.6	58.7	**59.9**	57.6	57.3	47.4	55.7	57.5	58.0	54.5	56.2
KindCLF	55.3	58.6	**59.3**	57.0	57.8	47.1	56.5	57.4	57.5	54.7	56.1
PreVPP	55.5	57.6	**59.6**	57.0	57.6	46.6	56.8	57.9	57.4	54.7	56.1
GoalPP	54.7	57.2	**58.4**	58.3	56.4	47.7	57.0	58.2	57.2	54.8	56.0
NPI	54.7	58.0	**60.5**	57.8	57.1	43.8	55.5	57.8	57.8	53.5	55.7
PathPP	55.4	58.6	58.8	56.7	**58.9**	45.0	54.9	56.3	57.4	53.8	55.6
AdvP	54.5	57.0	**58.7**	56.6	56.9	44.3	55.6	56.8	56.7	53.4	55.1
Deixis	54.1	56.7	**58.3**	56.5	56.6	46.0	54.7	55.7	56.4	53.2	54.8
VP	53.9	56.9	**58.3**	55.6	56.1	45.9	55.2	56.2	56.2	53.3	54.8
DirPP	54.0	57.7	**58.4**	55.0	55.6	46.6	54.6	55.9	56.3	53.2	54.7
VerbLE	54.4	56.6	**57.0**	56.2	56.2	44.6	55.4	55.6	56.0	53.0	54.5
MEINeg	53.3	56.4	**58.0**	56.5	56.0	44.4	55.5	55.8	56.0	52.9	54.5
ZHIQtf	54.0	55.9	**57.2**	56.5	55.3	44.9	55.7	55.8	55.9	52.9	54.4
CausCpl	53.5	55.5	**57.9**	54.9	56.5	43.7	54.6	56.0	55.5	52.7	54.1
SHICop	53.3	56.2	**57.2**	55.3	55.3	45.1	54.3	54.3	55.5	52.2	53.9
Cmpr	53.1	55.4	**57.4**	55.8	55.7	43.1	53.9	54.5	55.4	51.8	53.6
BUNeg	53.0	56.1	**57.3**	54.9	55.1	44.4	53.6	54.4	55.3	51.9	53.6
AdjP	52.7	55.8	**57.0**	55.0	55.0	42.7	53.3	54.6	55.1	51.4	53.3
SentIPP	53.5	55.0	**55.5**	54.2	53.1	41.7	52.9	52.9	54.5	50.2	52.4
Refl	51.2	53.3	**54.7**	53.2	52.7	42.4	52.4	53.4	53.1	50.2	51.7
ApprCLF	51.9	52.5	**55.4**	53.1	51.9	41.8	51.9	53.8	53.2	49.8	51.5
ConcCpl	49.8	51.3	**52.7**	49.8	51.0	39.2	51.4	51.2	50.9	48.2	49.6

Table 18: Performance in CHRF scores of each system on different grammatical features.

	Baidu	Niu	Google	DeepL	Ernie	Qwen	GPT	Claude	NMT-AVG	LLM-AVG	All-AVG
ZAIProg	83.1	83.3	**84.5**	82.8	84.0	79.0	83.3	83.1	83.4	82.3	82.9
SpanPP	82.9	83.0	**84.0**	82.7	83.7	78.8	83.3	83.5	83.2	82.3	82.7
ImpNeg	83.2	82.7	**85.1**	82.8	83.3	77.8	83.1	83.1	83.5	81.8	82.6
GUOPrf	82.2	82.3	**83.2**	81.8	**83.2**	78.5	83.1	83.0	82.4	81.9	82.2
StdCLF	82.1	82.4	**83.7**	81.8	83.6	78.1	82.5	82.4	82.5	81.7	82.1
Pass	82.3	82.5	**83.9**	82.3	82.4	77.1	82.4	83.1	82.8	81.2	82.0
BA	82.2	82.9	**83.8**	82.2	82.5	77.2	82.6	82.2	82.8	81.1	82.0
CondPP	82.0	83.0	**84.2**	81.6	82.8	77.4	82.0	82.4	82.7	81.2	81.9
TmpSCpl	82.1	82.3	**83.1**	82.0	82.6	77.8	82.5	82.4	82.4	81.3	81.9
VerbLE	81.9	82.1	**82.9**	82.2	82.5	77.8	82.4	82.2	82.3	81.2	81.8
MEINeg	81.6	81.5	**83.4**	81.8	82.9	77.0	82.5	82.6	82.1	81.2	81.7
NPI	81.9	82.2	**83.2**	81.8	82.3	76.9	81.8	82.3	82.3	80.8	81.5
SrcPP	81.9	81.9	**83.0**	80.9	82.3	77.2	82.3	82.6	81.9	81.1	81.5
CausCpl	81.5	81.1	**83.1**	80.8	82.6	77.0	82.9	82.9	81.6	81.3	81.5
Rel	81.4	81.8	**83.4**	81.5	82.3	77.3	81.6	82.4	82.0	80.9	81.5
ZHIQtf	81.6	81.6	**82.8**	81.6	82.6	76.5	82.6	82.3	81.9	81.0	81.4
RefPP	81.8	82.0	**82.4**	81.2	82.0	78.1	81.7	82.0	81.9	81.0	81.4
SpcPP	81.3	82.0	**83.0**	81.4	82.0	77.9	81.8	81.7	81.9	80.8	81.4
IndCLF	81.3	81.2	**82.9**	81.3	82.1	78.0	82.0	81.9	81.7	81.0	81.3
VP	81.3	81.6	**82.9**	81.0	82.2	77.2	81.9	82.2	81.7	80.9	81.3
AdvP	81.3	81.4	**82.7**	81.0	82.0	76.5	81.8	81.9	81.6	80.6	81.1
SHICop	80.8	81.1	**82.6**	80.5	81.9	77.1	81.9	81.6	81.2	80.6	80.9
Deixis	81.2	81.2	**82.5**	80.6	81.7	76.7	81.6	81.4	81.4	80.3	80.9
LocPP	80.8	82.0	**82.9**	81.1	81.6	76.2	81.2	81.2	81.7	80.0	80.9
Refl	80.8	81.0	**82.5**	80.9	81.6	76.4	81.6	81.9	81.3	80.4	80.8
PstVPP	81.2	80.7	**82.0**	80.6	81.4	76.6	**82.0**	81.9	81.1	80.5	80.8
NP	80.8	81.1	**82.4**	80.6	81.7	76.7	81.2	81.4	81.2	80.2	80.7
EvCLF	80.6	80.8	**82.1**	81.0	81.6	76.3	80.9	81.4	81.1	80.0	80.6
GoalPP	80.8	80.0	**81.6**	80.1	81.4	76.3	**81.6**	81.5	80.6	80.2	80.4
AdjP	80.2	80.6	**81.7**	80.3	81.2	75.4	81.0	81.2	80.7	79.7	80.2
Cmpr	80.3	80.4	**81.6**	80.2	81.4	75.6	80.9	81.0	80.6	79.7	80.2
PtcpPP	80.4	81.0	**81.7**	79.9	80.7	76.5	80.6	80.7	80.8	79.6	80.2
ConcCpl	80.3	80.5	**81.7**	80.1	81.1	75.1	81.4	80.9	80.7	79.6	80.1
SentIPP	80.5	80.5	**81.3**	79.9	80.6	75.4	80.6	80.3	80.6	79.2	79.9
DirPP	80.2	80.5	81.1	79.3	**81.2**	75.1	80.9	80.7	80.3	79.5	79.9
TopPP	80.1	80.2	**81.2**	79.9	80.6	74.9	80.6	80.5	80.3	79.2	79.8
Recp	80.0	80.0	**81.3**	79.8	80.6	74.8	80.3	80.5	80.3	79.0	79.7
LexNeg	79.3	79.5	**80.8**	80.0	**80.8**	76.4	79.9	80.4	79.9	79.4	79.6
ApprCLF	79.7	80.2	**81.5**	78.9	80.5	73.7	80.5	80.5	80.1	78.8	79.5
BUNeg	79.2	79.7	**80.4**	78.4	79.8	74.4	79.6	79.4	79.4	78.3	78.9
PreVPP	78.8	79.5	**81.0**	78.4	79.6	73.9	79.8	79.5	79.4	78.2	78.8
PathPP	78.4	79.1	**80.0**	78.3	79.5	73.7	78.7	79.7	79.0	77.9	78.4
KindCLF	77.8	78.2	**78.9**	77.3	78.3	73.9	78.4	78.5	78.0	77.3	77.7

Table 19: Performance in COMET scores of each system on different grammatical features.

	Baidu	Niu	Google	DeepL	Ernie	Qwen	GPT	Claude	NMT-AVG	LLM-AVG	All-AVG
ZAIProg	92.9	92.5	**93.7**	92.2	93.6	88.8	93.5	93.1	92.8	92.2	92.5
MEINeg	92.1	91.7	93.9	92.1	**94.0**	88.2	93.3	93.5	92.5	92.2	92.3
NPI	92.0	92.4	**93.0**	91.6	**93.0**	88.3	92.6	**93.0**	92.2	91.7	92.0
ImpNeg	91.9	91.0	**93.5**	92.1	92.8	87.4	93.2	92.7	92.1	91.5	91.8
Recp	92.0	91.0	**93.3**	91.2	92.8	87.7	92.6	92.5	91.9	91.4	91.6
Pass	92.0	90.8	**93.2**	91.2	92.6	86.9	92.6	92.8	91.8	91.2	91.5
ZHIQtf	91.3	90.8	**93.0**	91.2	92.5	86.6	92.4	92.2	91.6	90.9	91.2
SHICop	91.6	90.6	92.2	90.1	**92.4**	87.0	**92.4**	92.0	91.1	91.0	91.0
Refl	91.1	90.0	92.7	90.3	**92.8**	85.7	92.5	92.7	91.0	90.9	91.0
GUOPrf	91.0	90.2	92.1	90.9	91.9	86.9	**92.4**	91.6	91.0	90.7	90.9
StdCLF	91.4	90.9	**92.3**	89.8	**92.3**	86.0	**92.3**	91.8	91.1	90.6	90.8
Deixis	91.3	90.7	**92.7**	89.4	92.0	86.3	92.1	91.8	91.0	90.5	90.8
SpanPP	91.1	90.3	92.0	90.1	**92.6**	86.2	92.2	91.5	90.9	90.6	90.7
VP	90.8	90.3	**92.3**	90.7	91.6	86.5	91.7	91.6	91.0	90.3	90.7
CausCpl	91.4	89.4	92.6	88.6	92.4	84.4	92.7	**92.8**	90.5	90.6	90.5
CondPP	91.0	91.1	**92.8**	89.0	91.7	86.1	91.3	91.2	91.0	90.1	90.5
SrcPP	90.4	90.7	**92.4**	90.0	91.8	84.9	91.9	91.9	90.9	90.1	90.5
GoalPP	91.1	90.6	91.8	89.1	91.4	86.1	**92.0**	91.6	90.7	90.3	90.5
VerbLE	91.0	90.0	91.6	90.4	91.4	85.4	**91.8**	91.5	90.8	90.0	90.4
PstVPP	91.0	90.5	91.6	89.5	91.3	85.6	**91.7**	**91.7**	90.7	90.1	90.4
AdvP	90.7	89.9	**92.0**	89.6	91.5	85.4	91.6	91.6	90.6	90.0	90.3
IndCLF	90.3	90.0	**91.6**	89.2	91.3	86.8	91.1	90.6	90.3	89.9	90.1
TopPP	90.4	90.0	91.5	89.8	**91.6**	85.1	91.3	91.0	90.4	89.8	90.1
RefPP	90.6	90.0	**91.5**	89.6	91.2	86.2	90.9	90.8	90.4	89.8	90.1
ConcCpl	90.5	89.6	**92.1**	89.5	91.6	83.6	91.9	91.7	90.4	89.7	90.1
BA	90.4	90.4	**91.5**	89.9	90.7	84.3	90.6	90.6	90.6	89.1	89.8
BUNeg	89.9	89.1	90.9	89.2	**91.4**	84.5	90.9	90.6	89.8	89.3	89.6
DirPP	89.9	89.7	90.8	87.5	**90.9**	85.7	90.8	90.6	89.5	89.5	89.5
SpcPP	89.4	89.6	**91.0**	89.3	90.2	85.6	90.2	90.4	89.8	89.1	89.5
LocPP	89.5	89.8	**91.3**	89.1	90.9	83.4	90.8	90.9	89.9	89.0	89.4
EvCLF	89.5	89.2	**90.8**	89.6	90.3	85.0	89.9	**90.8**	89.8	89.0	89.4
NP	89.7	89.3	**90.9**	88.6	90.6	84.8	90.4	90.2	89.6	89.0	89.3
Cmpr	89.4	89.1	90.9	88.8	90.9	83.7	90.6	**91.0**	89.5	89.1	89.3
PreVPP	89.8	89.6	**91.4**	88.5	89.9	84.6	90.2	89.7	89.8	88.6	89.2
SentIPP	90.1	89.5	**90.6**	88.6	90.5	83.7	90.1	89.9	89.7	88.5	89.1
AdjP	89.4	88.9	**90.8**	88.4	90.7	83.5	90.5	90.2	89.4	88.7	89.0
TmpSCpl	89.7	89.3	**90.8**	88.5	90.3	83.0	90.0	90.2	89.6	88.4	89.0
Rel	89.5	88.4	**91.1**	88.0	90.4	82.7	90.2	90.4	89.2	88.4	88.9
PtcpPP	89.1	89.0	**90.7**	88.3	90.4	83.6	89.9	89.6	89.3	88.4	88.8
LexNeg	88.7	88.4	**90.2**	88.4	90.0	84.1	89.7	90.0	88.9	88.5	88.7
ApprCLF	89.1	88.2	90.5	87.8	**90.6**	82.8	90.1	90.0	88.9	88.4	88.6
KindCLF	89.0	88.4	**89.9**	88.0	89.7	83.8	89.4	89.0	88.8	88.0	88.4
PathPP	84.9	84.5	**86.4**	83.6	85.9	79.4	85.1	85.9	84.8	84.1	84.4

Table 20: Performance in XCOMET scores of each system on different grammatical features.

	Baidu	Niu	Google	DeepL	Ernie	Qwen	GPT	Claude	NMT-AVG	LLM-AVG	All-AVG
ZAIProg	74.3	73.8	74.7	73.8	75.3	68.8	**75.4**	74.7	74.2	73.5	73.9
SpanPP	74.8	73.8	74.9	73.6	75.1	68.1	**75.2**	74.8	74.3	73.3	73.8
CausCpl	74.6	73.3	75.6	73.6	75.0	65.4	75.5	**76.0**	74.3	73.0	73.6
TmpSCpl	74.7	73.8	74.3	73.1	74.7	67.5	74.8	**74.9**	74.0	73.0	73.5
MEINeg	74.5	73.2	75.1	74.1	75.1	65.5	**75.2**	74.9	74.2	72.7	73.4
VerbLE	74.3	73.2	74.4	73.7	74.6	66.7	**75.1**	74.8	73.9	72.8	73.4
BA	74.2	73.6	73.9	73.0	74.5	65.8	**75.0**	74.4	73.7	72.4	73.0
SpcPP	73.7	73.2	**74.3**	73.1	74.0	66.8	**74.3**	73.1	73.6	72.1	72.8
Recp	73.4	72.6	74.2	73.1	73.7	66.4	**74.5**	74.1	73.3	72.2	72.8
NPI	73.8	73.3	73.3	72.7	74.0	66.6	**74.5**	73.8	73.3	72.2	72.7
RefPP	73.7	72.4	73.5	72.3	73.8	67.3	**74.3**	73.8	73.0	72.3	72.7
AdvP	73.6	72.4	73.9	72.6	74.0	66.2	**74.4**	74.1	73.1	72.2	72.6
ConcCpl	73.4	72.2	74.4	72.7	74.5	64.2	**74.8**	74.6	73.2	72.0	72.6
StdCLF	74.1	72.2	73.4	72.7	74.1	65.5	**74.5**	73.6	73.1	71.9	72.5
VP	73.4	72.5	73.8	72.4	73.8	65.8	**74.0**	73.6	73.0	71.8	72.4
Refl	73.2	71.6	73.6	71.4	73.5	65.7	73.6	**74.1**	72.5	71.7	72.1
AdjP	73.2	72.4	73.2	72.3	73.6	64.2	**74.1**	73.5	72.8	71.3	72.1
PtcpPP	73.4	71.6	73.3	72.0	73.5	66.0	**73.7**	72.9	72.6	71.5	72.0
SHICop	73.5	71.6	73.2	71.6	73.9	64.8	**74.2**	72.9	72.5	71.4	72.0
ZHIQtf	72.7	71.7	73.1	72.8	73.5	64.7	**73.9**	73.3	72.6	71.3	71.9
SrcPP	73.1	71.9	73.0	71.3	73.3	65.6	**73.8**	73.5	72.3	71.5	71.9
Rel	73.4	71.3	72.9	71.5	73.5	65.0	**73.9**	73.2	72.3	71.4	71.8
Cmpr	72.7	71.6	73.1	72.0	73.4	63.6	73.8	**73.9**	72.3	71.2	71.8
LexNeg	72.4	71.0	72.6	72.0	73.1	66.3	**73.2**	73.0	72.0	71.4	71.7
GUOPrf	72.8	71.6	73.0	70.9	73.1	66.1	**73.3**	72.8	72.1	71.3	71.7
SentIPP	72.8	72.2	73.1	71.3	73.2	63.7	**73.5**	73.0	72.3	70.8	71.6
Pass	73.1	71.0	73.0	72.0	72.9	64.8	**73.2**	72.9	72.3	70.9	71.6
EvCLF	72.5	71.7	72.9	72.5	72.4	64.1	72.8	**73.0**	72.4	70.6	71.5
LocPP	72.7	71.8	72.2	72.1	72.8	64.6	**73.0**	72.7	72.2	70.8	71.5
NP	72.6	71.4	72.5	71.3	72.7	64.9	**73.2**	72.7	72.0	70.9	71.4
CondPP	**73.2**	71.3	71.8	70.7	**73.2**	65.6	73.0	72.4	71.8	71.1	71.4
ImpNeg	72.7	70.8	72.8	72.2	72.7	64.4	**73.5**	71.9	72.1	70.6	71.4
GoalPP	72.5	71.9	73.1	70.6	**73.2**	64.1	72.5	72.6	72.0	70.6	71.3
IndCLF	72.1	71.6	72.3	71.0	72.3	65.6	**73.0**	72.7	71.8	70.9	71.3
PstVPP	72.6	71.6	72.6	70.6	**73.3**	64.0	72.8	72.9	71.8	70.8	71.3
DirPP	71.6	71.4	72.1	69.8	73.3	63.5	**73.5**	**73.5**	71.2	71.0	71.1
Deixis	71.9	70.7	71.9	70.3	72.5	65.3	**72.7**	72.5	71.2	70.8	71.0
BUNeg	71.3	71.0	72.7	71.2	72.6	63.3	**73.0**	72.0	71.5	70.2	70.9
PreVPP	71.9	71.0	72.0	70.2	72.0	63.7	**72.1**	71.6	71.3	69.8	70.6
TopPP	70.6	69.5	70.9	70.3	71.1	62.1	**71.3**	70.8	70.3	68.8	69.6
KindCLF	70.4	69.0	70.6	69.6	70.9	63.5	**71.3**	70.7	69.9	69.1	69.5
ApprCLF	70.1	69.9	69.7	68.9	**70.2**	60.6	70.1	70.1	69.7	67.8	68.7
PathPP	70.1	68.2	68.9	68.4	69.3	62.9	69.9	**70.2**	68.9	68.1	68.5

Table 21: Performance in COMETKIWI-QE scores of each system on different grammatical features.

	Baidu	Niu	Google	DeepL	Ernie	Qwen	GPT	Claude	NMT-AVG	LLM-AVG	All-AVG
ZAIProg	95.3	94.5	94.9	94.4	95.6	91.6	**95.7**	95.3	94.8	94.5	94.7
MEINeg	94.9	94.2	95.3	94.0	**95.5**	90.8	95.4	**95.5**	94.6	94.3	94.4
NPI	95.1	94.6	94.6	93.5	**95.2**	91.3	**95.2**	95.0	94.4	94.2	94.3
ZHIQtf	95.0	94.3	94.6	94.2	95.0	90.7	**95.1**	94.5	94.5	93.8	94.2
Recp	94.5	93.9	94.9	94.0	94.5	91.7	**95.2**	94.5	94.3	94.0	94.1
VP	94.6	93.6	94.6	93.4	94.5	90.4	**94.7**	94.5	94.0	93.5	93.8
ConcCpl	94.6	93.4	94.7	93.7	94.3	88.3	**95.2**	94.5	94.1	93.1	93.6
Refl	94.4	92.8	94.5	92.7	94.1	90.0	94.5	**94.6**	93.6	93.3	93.5
SHICop	94.6	93.0	94.0	92.5	94.2	90.2	**95.0**	94.0	93.5	93.3	93.4
AdvP	94.2	92.9	94.0	92.5	93.9	89.6	**94.6**	94.1	93.4	93.1	93.2
StdCLF	**94.4**	93.2	93.5	92.7	93.9	89.6	94.3	93.7	93.5	92.9	93.2
GUOPrf	**94.0**	92.9	93.9	92.8	**94.0**	90.0	**94.0**	93.7	93.4	92.9	93.2
Deixis	94.1	92.9	93.8	92.0	94.0	89.8	**94.5**	94.0	93.2	93.1	93.1
CausCpl	94.3	92.0	94.4	91.5	94.0	88.0	**94.8**	**94.8**	93.1	92.9	93.0
VerbLE	93.5	92.7	93.7	92.9	93.3	89.5	**94.3**	93.7	93.2	92.7	93.0
SentIPP	93.8	92.8	**94.2**	92.1	93.8	89.2	93.8	93.3	93.2	92.5	92.9
Pass	**93.8**	92.0	93.7	92.4	93.4	90.1	93.3	93.7	93.0	92.6	92.8
RefPP	93.6	92.4	93.6	92.2	93.6	90.0	**93.8**	93.0	93.0	92.6	92.8
BUNeg	93.3	92.6	**93.7**	92.6	93.5	89.2	**93.7**	93.5	93.0	92.5	92.8
ImpNeg	93.9	92.2	93.1	92.9	93.5	89.0	**94.1**	93.2	93.0	92.5	92.7
SpanPP	93.4	92.1	93.7	92.0	93.7	89.7	**93.8**	93.2	92.8	92.6	92.7
SpcPP	93.2	92.8	**93.4**	92.0	93.1	89.2	**93.4**	93.2	92.8	92.2	92.5
BA	93.4	92.7	92.8	92.4	93.4	88.1	**93.8**	93.2	92.8	92.1	92.5
EvCLF	**93.3**	92.4	92.9	92.9	92.7	88.8	**93.3**	93.0	92.9	92.0	92.4
PstVPP	**94.0**	92.6	92.9	91.5	93.5	87.8	93.1	93.1	92.8	91.9	92.3
IndCLF	93.0	92.1	92.7	91.5	93.1	89.6	**93.5**	92.9	92.3	92.3	92.3
AdjP	93.3	91.7	93.2	91.6	93.2	88.2	**93.8**	93.0	92.4	92.0	92.3
GoalPP	**93.9**	92.6	93.1	90.9	93.4	87.7	93.0	92.8	92.6	91.7	92.2
DirPP	93.1	92.3	92.8	91.0	93.0	88.4	**93.7**	92.7	92.3	92.0	92.1
TopPP	92.7	92.1	92.9	92.0	92.8	89.0	**93.0**	92.5	92.4	91.8	92.1
TmpSCpl	**93.3**	92.1	92.7	91.1	92.9	86.9	**93.3**	93.0	92.3	91.5	91.9
Cmpr	92.6	91.3	92.9	91.2	93.2	87.5	**93.4**	93.1	92.0	91.8	91.9
NP	92.9	91.7	92.4	91.1	92.7	88.4	**93.1**	92.4	92.0	91.7	91.8
CondPP	93.1	91.5	92.4	90.4	92.8	88.5	**93.3**	92.2	91.8	91.7	91.8
SrcPP	92.2	91.4	92.3	91.2	**92.5**	88.1	**92.5**	92.0	91.8	91.3	91.5
PtcpPP	92.5	91.0	92.3	90.9	92.6	87.5	**93.0**	92.1	91.7	91.3	91.5
LexNeg	92.3	91.1	92.3	91.4	91.7	88.1	**92.4**	92.1	91.8	91.1	91.4
Rel	92.5	90.4	92.1	90.3	92.2	87.2	**93.3**	92.0	91.3	91.2	91.3
KindCLF	**92.2**	90.9	91.9	90.9	**92.2**	87.8	**92.2**	91.7	91.5	91.0	91.2
PreVPP	92.0	91.8	**92.2**	90.3	92.1	87.5	92.1	91.6	91.6	90.8	91.2
LocPP	91.9	91.4	91.5	91.4	91.8	86.6	**92.4**	92.1	91.6	90.7	91.1
ApprCLF	92.0	91.7	91.6	90.3	91.9	87.0	**92.2**	91.3	91.4	90.6	91.0
PathPP	**89.1**	86.4	87.2	86.6	87.5	84.5	88.3	88.8	87.3	87.3	87.3

Table 22: Performance in XCOMET-QE scores of each system on different grammatical features.